



Abstract. *As interdisciplinary and design-based learning approaches gain importance in teacher education, there is an increasing need for valid and reliable instruments to evaluate pre-service teachers' perceptions of STREAM education and related competencies. This study examined the psychometric properties of the Turkish adaptation of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers (STREAM-PICS) and its relationship with engineering design self-efficacy. A methodological and relational survey design was employed with 498 pre-service teachers in Türkiye. The adaptation process included translation, expert review, back-translation, and pilot-testing. Confirmatory factor analysis using the DWLS estimation method was performed due to the ordinal data structure and violation of multivariate normality assumptions; four items with low factor loadings and high error variances were removed. The final 26-item model confirmed a three-factor structure: perceptions of STREAM implementation, STREAM disciplines, and implementation challenges. Reliability analyses showed acceptable Cronbach's alpha, McDonald's omega, and composite reliability for all sub-dimensions. Correlation analyses revealed positive relationships between the first two STREAM dimensions and engineering design self-efficacy, whereas the implementation challenges dimension showed weak relationships. Significant differences were found by department, whereas no significant differences emerged by grade level.*

Keywords: *engineering design self-efficacy, measurement invariance, pre-service teachers, scale adaptation, STREAM education*

Halime Yerli, Derya Girgin
Çanakkale Onsekiz Mart University, Türkiye



PERCEPTIONS OF STREAM AND SELF-EFFICACY IN ENGINEERING DESIGN SKILLS AMONG PRE- SERVICE TEACHERS: SCALE ADAPTATION, VALIDITY- RELIABILITY, AND RELATIONAL EVIDENCE

**Halime Yerli,
Derya Girgin**

Introduction

In recent years, the growing emphasis on interdisciplinary learning, creativity, and problem-solving in education has increased interest in integrated approaches such as STEM, STEAM, and STREAM education. Today's education systems expect students not only to acquire discipline-specific knowledge but also to apply it in processes such as solving real-world problems, developing original products, integrating diverse perspectives, and producing sustainable solutions (Johnston et al., 2022; Sun & Zhong, 2024). This expectation has brought the STEM approach, which integrates science, technology, engineering, and mathematics, to the centre of educational research; later, it laid the groundwork for the development of the STEAM approach, which emphasizes the creative and aesthetic representation power of art, and the STREAM approach, which adds the literacy/reading dimension to the interdisciplinary meaning-making process. In this developmental line, integrative reviews on STEAM education have emphasized that interdisciplinary learning should be understood not only as the addition of art to STEM but also as a pedagogical framework that connects creativity, inquiry, design, and meaning-making processes (Ng et al., 2022). STREAM education is a contemporary educational approach that aims to develop students' skills in scientific inquiry, technological production, literacy, engineering design, creative expression, and mathematical reasoning through holistic learning experiences by bringing together the fields of Science, Technology, Reading, Engineering, Arts, and Mathematics (Pakalnina & Grava, 2025; Phang et al., 2023; Sun & Zhong, 2024). In this respect, STREAM is not merely about bringing disciplines together; it offers an integrated pedagogical framework that supports students in analysing problem situations from multiple dimensions, justifying their solutions, utilizing design and production processes, and creating meaning through different forms of representation. Similarly,



STREAM education has been discussed as a multidimensional educational approach that involves prospects, priorities, processes, and implementation-related problems rather than a simple combination of disciplinary fields (Belbase et al., 2022). Although STEM (Science, Technology, Engineering, and Mathematics), STEAM (STEM + Arts), and STREAM (STEM + Reading/wRiting) share a common interdisciplinary foundation, they differ in their curricular emphasis and scope. While STEM focuses on technical competencies and STEAM incorporates creative and artistic dimensions, STREAM uniquely integrates literacy components - reading and writing - into the interdisciplinary framework, thereby fostering language-based reasoning alongside scientific and engineering practices (Sun & Zhong, 2024). The present study specifically focuses on STREAM as a distinct and emerging approach, rather than treating these frameworks interchangeably.

One of the main reasons for the increasing importance of the STREAM approach is its close relationship with 21st-century skills and interdisciplinary learning processes. Skills such as creative thinking, problem-solving, collaboration, communication, decision-making, critical thinking, and innovative product development develop more robustly at the intersection of disciplines rather than within the boundaries of a single discipline. Current studies in the field of science education also show that integrated STEM, problem-based learning, and design-based applications can have positive effects on students' conceptual understanding, scientific creativity, problem-solving, motivation, and entrepreneurial creative thinking skills (Arifin & Siew, 2023; Dedetürk et al., 2021; Ghazali et al., 2024; Khamhaengpol et al., 2021). The theoretical background of STREAM education can be associated with social constructivist learning, design-based learning, and self-efficacy theory. The social constructivist approach emphasizes that knowledge is constructed through social interaction, language, cultural tools, and collaboration, while design-based learning enables students to identify criteria and constraints, generate alternative solutions, develop prototypes, and evaluate their solutions in response to open-ended problem situations. In this regard, the engineering design process is one of the central components of the STREAM approach (Capobianco & Radloff, 2022; Mumba et al., 2023; Ojeogwu & Mumba, 2025).

In terms of teacher education, the transfer of the STREAM approach to classroom environments largely depends on the knowledge, perception, experience, and self-efficacy of teachers and pre-service teachers regarding this approach (Alanazi & Al-Zahrani, 2025; Subramaniam et al., 2023). Teachers are the key actors who design interdisciplinary learning environments, guide the learning process, and support students' design-production processes. In early childhood education contexts, teachers' beliefs about STEAM education also play a decisive role in how integrated interdisciplinary practices are interpreted and implemented in classroom environments (Alghamdi, 2023). Therefore, how pre-service teachers perceive STREAM practices, the extent to which they value STREAM disciplines, and how they evaluate the difficulties in the implementation process are crucial in determining the likelihood of them using this approach in their classrooms in the future. Phang et al. (2023) have shown that while teachers have positive perceptions of STREAM education, they face difficulties in terms of teacher expertise, professional training opportunities, resource shortages, and implementation confidence. Pakalnina and Grava (2025) also stated that for STREAM practices to be effective, teachers need to rethink the pedagogical process, be supported with appropriate methods and materials, and have their professional development opportunities strengthened.

In this context, the self-efficacy theory provides an important explanatory framework. According to Bandura (1997), self-efficacy is the belief in one's ability to successfully perform a specific task, and this belief affects the individual's effort, determination, resilience in the face of challenges, and performance. In the context of STEM and engineering education, teacher self-efficacy is related to their capacities to plan interdisciplinary activities, implement design-based teaching, support students' problem-solving processes, and cope with uncertainties in the teaching process. Chen et al. (2021) reported that pre-service teachers' STEM self-efficacy is related to their pedagogical beliefs and professional development needs; Yang et al. (2023) found that STEM teaching self-efficacy exhibits a two-dimensional structure of pedagogical self-efficacy and content self-efficacy. Additionally, the inclusion of the engineering design process in STEM project-based learning can support pre-service teachers' engineering design thinking, their ability to design STEM lessons, and their self-efficacy in science-engineering teaching (Capobianco et al., 2022; Hammack et al., 2024; Laksmiwati et al., 2024; Lin et al., 2021; Mumba et al., 2023).

In the context of Türkiye, engineering design skills are gaining increasing importance in science curricula in terms of interdisciplinary thinking, innovative product development, problem-solving, and productivity. The Engineering Design Skills Self-Efficacy Perception Scale, developed by Kahraman et al. (2024), allows for a multidimensional assessment of pre-service teachers' self-efficacy perceptions regarding engineering design skills. However, the relationships between engineering design self-efficacy and pre-service teachers' perceptions of STREAM implementations, their evaluation methods of STREAM disciplines, and the challenges of STREAM implementation have not yet been sufficiently clarified.



Three main gaps are noteworthy in the literature. Despite the increase in STEM/STEAM-focused studies, there are limited valid and reliable tools for measuring the STREAM approach in the context of pre-service teachers. Current studies mostly focus on narrower constructs such as STEM teaching self-efficacy, STEM attitude, STEM pedagogical belief, or engineering design self-efficacy (Chen et al., 2021; Hammack et al., 2024; Menon et al., 2025; Yang et al., 2023). In addition, in scale adaptation studies, it is necessary to present multiple pieces of evidence together, such as construct validity, reliability, convergent and discriminant validity, criterion-related validity, and measurement invariance, rather than relying solely on translation and internal consistency evidence (International Test Commission [ITC], 2017). Another gap concerns the lack of interpretable cut-off points for subscale scores that can support needs assessment, program evaluation, and professional development planning in teacher education. These gaps guided the research questions addressed in this study.

The unique value of this study stems from its comprehensive evaluation of the Turkish adaptation of the STREAM-PICS and its examination of the relationship between STREAM perceptions and engineering design self-efficacy among pre-service teachers. The study provides comprehensive psychometric evidence and empirical support for using the instrument in Turkish teacher education.

Research Aim and Research Questions

This study aimed to examine the psychometric properties of the Turkish version of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers and to evaluate the interpretability of its subscale scores. Specifically, the study examined the factor structure, reliability, validity, measurement invariance across gender, empirical cut-off scores, and the relationships between STREAM perceptions and engineering design self-efficacy. Differences according to class level and academic department were also examined.

1. Is the three-factor structure of the Turkish form of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers validated in a sample of pre-service teachers?
2. Do the standardized factor loadings and error variances of the items in the Turkish version of the scale exhibit acceptable psychometric properties at the item level?
3. Do the subscale scores obtained from the Turkish form of the STREAM scale provide sufficient evidence of reliability in terms of Cronbach's alpha, McDonald's omega, and composite reliability coefficients?
4. What is the relationship between the sub-dimensions of the STREAM scale and the total score and sub-dimensions of the Engineering Design Skills Self-Efficacy Perception Scale?
5. To what extent does the Turkish form of the STREAM scale provide evidence of convergent and discriminant validity in terms of AVE, CR, and HTMT values?
6. What empirical score classifications can be derived for interpreting the subscale scores of the STREAM scale?
7. Does the Turkish version of the STREAM scale provide configural, metric, scalar, and strict measurement invariance across gender?
8. Do the perceptions of pre-service teachers regarding STREAM implementations, their perceptions of STREAM disciplines, and their perceptions of the difficulties in implementing STREAM education show significant differences according to grade level?
9. Do the perceptions of pre-service teachers regarding STREAM implementations, their perceptions of STREAM disciplines, and their perceptions of the difficulties in implementing STREAM education show significant differences based on the department they are studying in?

Research Methodology

Design

This quantitative study combined methodological and relational survey designs. The methodological component evaluated the Turkish adaptation of the STREAM-PICS through construct validity, internal consistency, convergent and discriminant validity, criterion-related validity, gender-based measurement invariance, and score interpretability. The relational component examined the associations between STREAM-PICS subscale scores and engineering design self-efficacy. The study followed contemporary test-adaptation standards, which require evidence beyond linguistic equivalence, including structural validity, reliability, relations with external variables, and group comparability (American Educational Research Association et al., 2014; DeVellis, 2017; ITC, 2017).

Data were collected cross-sectionally during the 2025–2026 academic year. Preparatory procedures, including instrument selection, permission, translation, expert review, pilot testing, and ethical approval, were completed in the fall semester, and the finalized instruments were administered in the spring semester. Because the factor structure was theoretically specified, confirmatory factor analysis was used to evaluate the fit between the proposed model and the observed data (Brown, 2015; Creswell & Creswell, 2018; Kline, 2016).

Participants

Participants were selected through convenience sampling from education faculty programs accessible to the researchers, an approach commonly used when probability sampling is not feasible in scale-adaptation research (Büyükoztürk et al., 2019). Data were initially collected from 508 pre-service teachers. Ten cases identified as multivariate outliers through Mahalanobis distance were excluded, resulting in a final sample of 498 participants.

The final sample exceeded general recommendations for factor analysis and confirmatory factor analysis. The two instruments contained 61 items in total, corresponding to approximately 8.2 participants per item, which was within the recommended range of 5–10 participants per item (Hair et al., 2019; Tabachnick & Fidell, 2019). It also exceeded the minimum of 200 cases recommended for CFA models of moderate complexity (Wolf et al., 2013). A power analysis based on $\alpha = .05$, $f^2 = .15$, and .80 power indicated a minimum sample of 181 participants (Cohen, 1988; Faul et al., 2009).

Descriptive results regarding the sociodemographic information of the participants are provided in Table 1.

Table 1
Sociodemographic Information of Pre-Service Teachers

Variable	Category	<i>f</i>	%
Gender	Female	409	82.3
	Male	88	17.7
	Missing data	1	-
Grade level	1st year	128	25.7
	2nd year	116	23.3
	3rd year	144	28.9
	4th year	110	22.1
Previous participation in STREAM-related training	Yes	77	15.6
	No	418	84.4
	Missing data	3	-
Department	Primary Mathematics Education	116	23.3
	Chemistry Education	33	6.6
	Science Education	58	11.6
	Primary School Education	126	25.3
	Turkish Language Education	133	26.7
	Art Education	32	6.4
Taking courses related to STEM/STREAM/engineering design in the undergraduate program	Yes	40	8.1
	No	453	91.9
	Missing data	5	-
Experience in planning STEM activities within workshops/training/projects	Yes	43	8.7
	No	450	91.3
	Missing data	5	-

Variable	Category	f	%
Conducting design/production activities with students during school experience/teaching practicum courses	Yes	45	9.1
	No	449	90.9
	Missing data	4	-

Note. f = frequency.

As shown in Table 1, participants were distributed relatively evenly across grade levels, although women constituted 82.3% of the sample. Most participants had not received prior STREAM-related training (84.4%), taken relevant undergraduate courses (91.9%), planned STEM activities in workshops or projects (91.3%), or conducted design and production activities during teaching practice (90.9%). Thus, the sample largely represented pre-service teachers with limited prior exposure to STREAM and engineering design.

Instruments

Data were collected using a Personal Information Form, the Turkish STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers (STREAM-PICS), and the Engineering Design Skills Self-Efficacy Perception Scale. Engineering design self-efficacy was used as an external criterion because of its theoretical relationship with the design-based nature of STREAM education (DeVellis, 2017).

Personal Information Form

The form collected information on gender, grade level, academic department, previous STREAM-related training, relevant undergraduate coursework, experience in planning STEM activities, and design or production activities conducted during teaching practice.

Turkish Form of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers (STREAM-PICS)

The original STREAM-PICS was used by Phang et al. (2023) to assess perceptions of STREAM implementation, the educational value of STREAM disciplines, and challenges encountered during implementation. The original instrument comprised 30 items rated on a four-point Likert scale.

The instrument was adapted into Turkish in accordance with the ITC (2017) guidelines. Responses ranged from 1 (*strongly disagree*) to 4 (*strongly agree*). Higher scores indicated more positive perceptions of STREAM implementation and disciplines in the first two subscales and greater perceived implementation challenges in the third subscale. Following the adaptation and validation procedures, the Turkish form comprised 26 items across three factors. Detailed validity and reliability findings are reported in the Results section.

Engineering Design Skills Self-Efficacy Perception Scale

The Engineering Design Skills Self-Efficacy Perception Scale was developed by Kahraman et al. (2024) to assess pre-service teachers' perceived competence in engineering design. The instrument contains 35 positively worded items across five subscales: part-whole focus, product development, multidisciplinary collaboration, design creation, and generating original solutions and decision-making. Items are rated on a four-point Likert scale, and higher scores indicate greater engineering design self-efficacy.

In the present sample, internal consistency was high for the total score ($\omega = .964$, $\alpha = .954$), while subscale coefficients ranged from .843 to .911 for omega and from .845 to .911 for alpha. Confirmatory factor analysis supported the five-factor structure, $\chi^2/df = 1.948$, CFI = .978, NNFI = .977, NFI = .956, RMSEA = .045, 90% CI [.041, .049], and SRMR = .080.

Scale Adaptation Process

Permission to adapt the STREAM-PICS into Turkish was obtained from the original authors. Eight language experts contributed to the forward translation. The Turkish form was reviewed by four subject-matter experts, eight language experts, and two measurement and evaluation experts in terms of linguistic clarity, conceptual equivalence, cultural appropriateness, and suitability for pre-service teachers. Revisions were based on qualitative expert feedback; a content validity index was not calculated. The revised form was back-translated into English and compared with the original instrument to identify possible discrepancies. Pilot testing and cognitive interviews with pre-service teachers who had characteristics similar to those of the target sample were then used to evaluate item clarity, response options, and contextual relevance. Minor wording revisions were made before the application form was finalized. These procedures followed the ITC (2017) guidelines, which emphasize linguistic, conceptual, and cultural equivalence in test adaptation.

Following administration, construct validity was examined through confirmatory factor analysis. Items with low factor loadings and high error variances were evaluated jointly in terms of statistical performance, theoretical representativeness, content validity, and expert judgment. Items whose removal did not narrow the theoretical coverage of the construct were excluded, and the final model was retested (DeVellis, 2017; Kline, 2016; Worthington & Whittaker, 2006).

Data Collection Process and Ethical Principles

Ethical approval for the study was obtained from the Social and Humanities Ethics Committee of Çanakkale Onsekiz Mart University (Approval No. E-80110942-050.99-2500208162 – 2025-YÖNP-0709 – 55/72; Date: 22 August 2025), and institutional permission was secured before data collection. Participants received information about the study's purpose, voluntary participation, withdrawal rights, anonymity, confidentiality, and scientific use of the data. Completion of the form was regarded as informed consent. The Personal Information Form and both scales were administered together and required approximately 10–12 minutes to complete. No identifying information was collected, and the data were analysed anonymously in accordance with established ethical principles for research involving human participants (American Psychological Association [APA], 2017).

Data Analysis

The dataset was screened for missing values, univariate and multivariate outliers, and distributional assumptions. No item-level data were missing. Cases identified as multivariate outliers through Mahalanobis distance were excluded, and the analyses were conducted with 498 participants. Multivariate normality was assessed using Mardia's test. Because the assumption of multivariate normality was not met and the STREAM-PICS items were measured on a four-point ordinal scale, confirmatory factor analysis was conducted using diagonally weighted least squares estimation (Li, 2016; Rhemtulla et al., 2012; Tabachnick & Fidell, 2019).

Frequencies and percentages were calculated for participant characteristics. The hypothesized three-factor structure of the Turkish STREAM-PICS was evaluated through confirmatory factor analysis. Model fit was assessed using χ^2/df , RMSEA, CFI, TLI/NNFI, NFI, GFI, and SRMR. Values above .90 for incremental fit indices and below .08 for RMSEA and SRMR were considered acceptable, while values above .95 and below .05, respectively, indicated good fit (Brown, 2015; Hu & Bentler, 1999; Kline, 2016; Schermelleh-Engel et al., 2003). Item retention decisions were based jointly on standardized factor loadings, error variances, theoretical relevance, content coverage, and expert judgment rather than statistical criteria alone (Brown, 2015; DeVellis, 2017).

Internal consistency was evaluated using Cronbach's alpha, McDonald's omega, and composite reliability, with coefficients of .70 or higher considered acceptable. Omega and composite reliability were included because alpha alone may be affected by violations of tau-equivalence (Dunn et al., 2014; McNeish, 2018; Nunnally & Bernstein, 1994).

Criterion-related validity was examined through Pearson correlations between the STREAM-PICS subscales and the total and subscale scores of the Engineering Design Skills Self-Efficacy Perception Scale (DeVellis, 2017). Convergent validity was evaluated using average variance extracted (AVE), interpreted together with composite reliability when AVE was below .50 (Fornell & Larcker, 1981; Hair et al., 2019). Discriminant validity was assessed using heterotrait–monotrait ratios, with values below .85 considered supportive of factor distinctiveness (Henseler et al., 2015).

Gender-based measurement invariance was tested through configural, metric, scalar, and strict models. Model comparisons were based on changes in CFI, RMSEA, and SRMR because chi-square difference tests may be overly sensitive in large samples. Changes of no more than .01 in CFI, .015 in RMSEA, and .030 in SRMR were used as the principal criteria for invariance, while the more liberal Δ CFI criterion of .02 was considered for complex group comparisons (Chen, 2007; Cheung & Rensvold, 2002; Meredith, 1993; Putnick & Bornstein, 2016; Rutkowski & Svetina, 2014).

Two-step cluster analysis was conducted separately for each STREAM-PICS subscale to derive sample-based score classifications. Cluster quality and separation were evaluated using the average Silhouette coefficient. These classifications were treated as provisional empirical cut-off points rather than expert-based standards (AERA et al., 2014; Kaufman & Rousseeuw, 2009).

One-way ANOVA was used to examine differences in subscale scores by grade level and academic department. Homogeneity of variances was assessed using Levene's test; significant effects were followed by post hoc comparisons, and eta squared was reported as the effect-size estimate (Cohen, 1988; Lakens, 2013). Because the STREAM-PICS represents three related but distinct dimensions, subscale scores were used for interpretation; the total score was reported only for descriptive purposes.

Research Results

Within the scope of the research, confirmatory factor analysis (CFA) was conducted to obtain evidence of construct validity for the adapted measurement tool. When the initial CFA results were examined, it was determined that the standardized factor loadings of the four items in the scale fell below the acceptable lower limit and that the error variances were quite high. Accordingly, for item1, $\lambda = .210$, $\varepsilon = .956$; for item10, $\lambda = .226$, $\varepsilon = .949$; for item26 $\lambda = .203$, $\varepsilon = .957$ and for item30 $\lambda = .206$, $\varepsilon = .958$ were calculated. Therefore, these items have been examined not only based on statistical justifications but also in accordance with the evaluations made by field experts and measurement-evaluation specialists regarding content, scope, and measurement appropriateness.

As a result of expert opinions, it was evaluated that the removal of Items 1, 10, 26, and 30 would not compromise the theoretical structure or content validity of the scale. Accordingly, these items were removed, and the CFA model was re-estimated using the remaining items. The item removal process was guided by both statistical evidence and expert judgment to preserve the theoretical integrity of the scale. The fit indices obtained from the CFA are presented in Table 2.

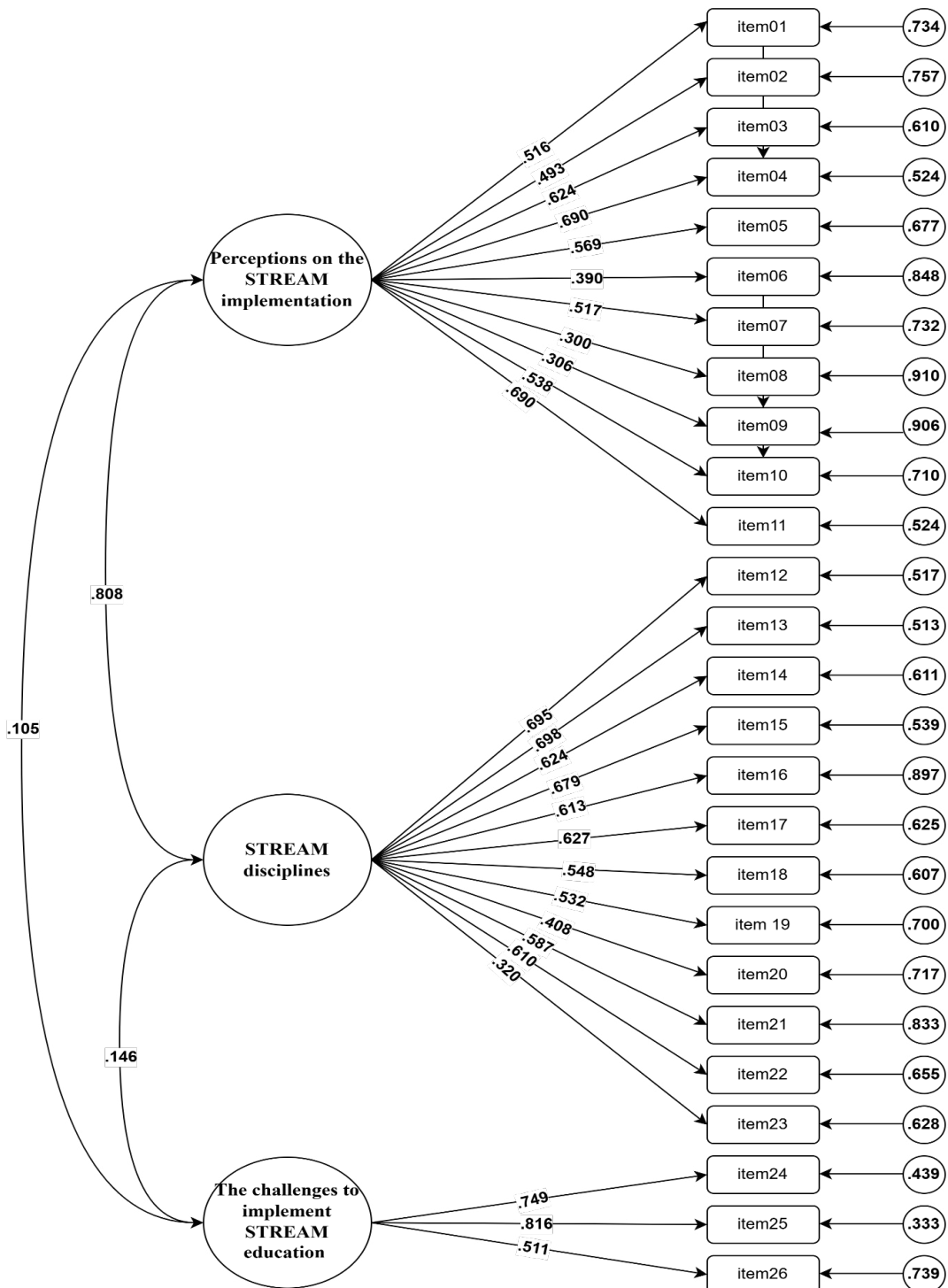
Table 2

Estimated Values and Threshold Values of Fit Indices for the STREAM Scale

Fit Indices	Good Fit	Acceptable Fit	Estimated Values
χ^2/df	$0 \leq \chi^2/df < 2$	$2 \leq \chi^2/df \leq 5$	1.998
RMSEA(%90 CI)	$0 \leq RMSEA < 0.05$	$0.05 \leq RMSEA \leq 0.10$	0.045 (0.040-0.050)
NNFI	$0.95 \leq NNFI \leq 1.00$	$0.90 \leq NNFI < 0.95$	0.962
NFI	$0.95 \leq NFI \leq 1.00$	$0.90 \leq NFI < 0.95$	0.933
CFI	$0.95 \leq CFI \leq 1.00$	$0.90 \leq CFI < 0.95$	0.965
GFI	$0.95 \leq GFI \leq 1.00$	$0.90 \leq GFI < 0.95$	0.998
SRMR	$0 \leq SRMR < 0.05$	$0.05 \leq SRMR \leq 0.10$	0.065

Note. RMSEA = Root Mean Square Error of Approximation; CI = Confidence Interval; NNFI = Non-Normed Fit Index; NFI = Normed Fit Index; CFI = Comparative Fit Index; GFI = Goodness of Fit Index; SRMR = Standardized Root Mean Square Residual. Adapted from Browne and Cudeck (1992).

The findings presented in Table 2 indicate that the tested measurement model showed an acceptable to good fit. The χ^2/df ratio calculated for the model was 1.998, which falls within the good fit range (MacCallum et al., 1996). Additionally, the RMSEA value of .045 indicates that the model also exhibited a high level of fit in terms of error sensitivity. The CFI, NNFI/TLI, and SRMR indices also met the recommended threshold values and confirmed the overall fit of the model (Hu & Bentler, 1999). These findings indicate that the three-factor structure of the scale was supported in the sample of pre-service teachers. The measurement model related to the tested model is provided in Figure 1.

Figure 1*The Tested Measurement Model*

Note. The figure presents the standardized factor loadings and error variances of the retested measurement model after the item-removal process. In the first stage of the adaptation process, Items 1, 10, 26, and 30 were removed from the measurement tool due to low factor loadings and high error variances, taking expert opinions into consideration. Following this procedure, the remaining items in the scale were renumbered sequentially, and the final confirmatory factor analysis (CFA) model was reported based on the revised item numbering. Therefore, the item numbers presented in the figure represent the current item order after the item-removal process.



When examining the item factor loadings of the validated model, the factor loadings ranged from $\lambda = .300$ to $\lambda = .816$, and the error variance values ranged from $\varepsilon = .333$ to $\varepsilon = .910$.

Most items demonstrated acceptable psychometric properties. Two items showed borderline factor loadings ($\lambda = .300$ and $\lambda = .306$) with relatively high error variances ($\varepsilon = .910$ and $\varepsilon = .906$).

The two items were retained because expert evaluations indicated that they contributed to the theoretical coverage of the construct. However, given their borderline factor loadings and relatively high error variances, these items should be re-examined and, if necessary, revised in future validation studies. After the validity evidence of the measurement tool was provided, the reliability evidence was examined and presented in Table 3.

Table 3
Reliability Coefficients for the Scale and Its Sub-dimensions

Scale / Sub-dimensions	Number of Items	McDonald's ω	Cronbach's α	Composite Reliability (CR)
Perceptions on the STREAM Implementation	11	.746	.800	.799
STREAM Disciplines	12	.824	.854	.869
The Challenges to Implement STREAM Education	3	.745	.725	.740
Scale Total	26	.879	.877	.930

Note. CR = Composite Reliability. The coefficients related to the total scale score are reported for descriptive purposes only. Due to the multidimensional structure of the scale, the correlation patterns among the sub-dimensions, and particularly the distinct nature of the “Challenges in Implementing STREAM Education” dimension compared to the other dimensions, sub-dimension scores rather than the total score were used in the interpretation of the scale.

Table 3 presents McDonald's omega, Cronbach's alpha, and composite reliability coefficients for the scale and its sub-dimensions. All reliability coefficients exceeded the recommended threshold of .70, indicating acceptable internal consistency for each subscale (Nunnally & Bernstein, 1994).

Convergent and Discriminant Validity

After providing evidence of construct validity for the measurements obtained from the adapted Stream scale, correlations with the Engineering Design Skills Self-Efficacy Perception Scale and its sub-dimensions, as well as AVE and HTMT ratios, were examined to provide evidence of criterion validity. The obtained findings are presented in tables 4 and 5.

Table 4
Correlations Between the STREAM Scale and Criterion Variables

Variables	1	2	3	4	5	6	7	8	9
1. Perceptions on the STREAM Implementation	—								
2. STREAM Disciplines	.628**	—							
3. The Challenges to Implement STREAM Education	.057	.110*	—						
4. STREAM Total Scale	.720**	.759**	.646**	—					
5. Part-Whole Focus	.402**	.288**	.062	.327**	—				
6. Product Development	.404**	.408**	.105*	.405**	.738**	—			
7. Multidisciplinary Collaboration	.371**	.394**	.107*	.387**	.594**	.703**	—		

Variables	1	2	3	4	5	6	7	8	9
8. Design Creation	.294**	.233**	.001	.224**	.527**	.515**	.442**	—	
9. Generating Original Solutions and Decision Making	.339**	.363**	.054	.330**	.603**	.682**	.614**	.527**	—
10. Engineering Design Skills Self-Efficacy Perception Scale	.442**	.402**	.079	.403**	.872**	.894**	.785**	.732**	.815**

Note. $N = 498$. Pearson correlation coefficients have been reported. * $p < .05$, ** $p < .01$.

Table 4 presents the Pearson correlation coefficients between the STREAM scale and the Engineering Design Skills Self-Efficacy Perception Scale and its sub-dimensions. Perceptions of STREAM implementation and STREAM disciplines were moderately and positively correlated ($r = .628, p < .01$). However, the relationships of the challenges to implement the STREAM education sub-dimension with the other two STREAM sub-dimensions are very weak. The relationship of this sub-dimension with Perceptions on the STREAM implementation is not significant ($r = .057, p > .05$); the relationship with STREAM disciplines is positive but at a very low level ($r = .110, p < .05$).

When examining the relationships between the sub-dimensions of the STREAM scale and the sub-dimensions of the Engineering Design Skills Self-Efficacy Perception Scale, it has been determined that the sub-dimension of Perceptions on the STREAM implementation shows positive and low to moderate significant relationships with the sub-dimensions of engineering design skills. This dimension's Part-whole-focus ($r = .402, p < .01$), Product development ($r = .404, p < .01$), Multidisciplinary collaboration ($r = .371, p < .01$), Design creation ($r = .294, p < .01$), and Generating original solutions and decision-making ($r = .339, p < .01$) sub-dimensions have significant relationships. Similarly, the STREAM disciplines sub-dimension also shows positive and significant relationships with the engineering design skills sub-dimensions. These relationships are at the levels of $r = .288$ for Part-whole focus, $r = .408$ for Product development, $r = .394$ for Multidisciplinary collaboration, $r = .233$ for Design creation, and $r = .363$ for Generating original solutions and decision-making ($p < .01$). In contrast, the relationships between the sub-dimension of The challenges to implement STREAM education and the sub-dimensions of engineering design skills are generally very low; only statistically significant but very weak relationships were found with Product development ($r = .105, p < .05$) and Multidisciplinary collaboration ($r = .107, p < .05$). The relationships with part-whole focus, design creation, generating original solutions and decision-making are not significant.

The relationships between the total score of the Engineering Design Skills Self-Efficacy Perception Scale and the STREAM sub-dimensions also support this pattern. The total score of engineering design skills shows a positive and moderate relationship with Perceptions on the STREAM implementation ($r = .442, p < .01$), it shows a positive and moderately strong relationship with STREAM disciplines ($r = .402, p < .01$). However, the relationship between the challenges to implement the STREAM education sub-dimension and the total score of engineering design skills is not significant ($r = .079, p > .05$). This situation reveals that the sub-dimensions of the STREAM scale do not show relationships at the same level and in a similar direction with engineering design skills.

Table 5
AVE and HTMT Values for the STREAM Scale

Factors	AVE	Factor 1	Factor 2	Factor 3
Factor 1: Perceptions on the STREAM Implementation	.259	—		
Factor 2: STREAM Disciplines	.320	.765	—	
Factor 3: The Challenges to Implement STREAM Education	.498	.133	.134	—

Note. AVE = Average Variance Extracted; HTMT = Heterotrait–Monotrait Ratio. The values below the diagonal represent the HTMT ratios between the factors.

Evidence of convergent and discriminant validity was evaluated using AVE, HTMT ratios, composite reliability and inter-factor correlations. The AVE values were .259 for Factor 1, .320 for Factor 2, and .498 for Factor 3, whereas composite reliability coefficients exceeded .70 for all factors. However, the acceptable levels of composite reliability

coefficients and the support of HTMT values for discriminant validity provide partial psychometric evidence for the usability of the scale at the sub-dimension level. Therefore, when evaluating the psychometric properties of the scale, it should be noted that the evidence of convergent validity is limited rather than strong, whereas the evidence of construct reliability and discriminant validity is more supportive.

The findings obtained in terms of discriminant validity provide stronger evidence. When examining the inter-factor correlations, it is observed that there is a positive and moderately significant relationship between Factor 1 and Factor 2, $r = .628$, $p < .01$. However, the relationship between Factor 3 and Factor 1 is not significant, $r = .057$, $p > .05$; Its relationship with Factor 2 is positive but quite low, $r = .110$, $p < .05$.

HTMT ratios also supported discriminant validity. The HTMT value between Factor 1 and Factor 2 was .765, the HTMT value between Factor 1 and Factor 3 was .133, and the HTMT value between Factor 2 and Factor 3 was .134. These values were below the .85 threshold, indicating discriminant validity among the factors. Especially the very low HTMT values between Factor 1–Factor 3 and Factor 2–Factor 3 indicate that the third factor is conceptually and statistically distinct from the other factors. The HTMT value of .765 between Factor 1 and Factor 2 indicates that these two dimensions are related but do not overlap to a degree that would threaten their discriminant validity.

Criterion-related validity findings were consistent with the correlation analyses. The first two STREAM sub-dimensions showed moderate positive relationships with engineering design self-efficacy, whereas the “Challenges to Implement STREAM Education” showed weak or non-significant relationships.

Determination of Empirical Cut-off Points for Subscale Scores

To enhance the interpretability of the scores obtained from the scale and to determine empirical threshold values for STREAM levels, a two-stage cluster analysis was conducted. Due to the low correlations between the sub-dimensions, the sub-dimension scores were used instead of the total score of the scale in the empirical classification process. Therefore, the obtained levels were interpreted not as a classical standard-setting process based on expert judgment, but as empirical cut-off points based on the score distributions obtained from the current sample. In this context, a two-stage cluster analysis was conducted based on the sub-dimensions.

As a result of the analysis, it was determined that the first factor showed a three-cluster structure, while the second and third factors showed a two-cluster structure. The clustering quality for F1 is .645, the average Silhouette coefficient is .524; the clustering quality for F2 is .696, the average Silhouette coefficient is .574, and for F3, the clustering quality is .694, the average Silhouette coefficient was obtained as .571. Considering that the Silhouette coefficient takes values between -1 and $+1$, and values approaching $+1$ indicate a more consistent clustering structure, it is seen that the obtained value is at an acceptable level. Additionally, clustering quality values above .50 indicate that the data are sufficiently separated into clusters (Dinh et al., 2019).

In the interpretation of the scores related to the F1 dimension, a three-cluster structure was taken as the basis, and two cutoff points were determined accordingly. Accordingly, scores in the range of 1.00–2.80 from the F1 dimension were evaluated as low level, scores in the range of 2.81–3.30 as medium level, and scores in the range of 3.31–4.00 as high level. According to this classification, high scores indicate that participants have a more positive/higher perception of STREAM implementations, while low scores show that this perception is at a more limited level. In the interpretation of the scores related to the F2 dimension, a two-cluster structure was taken as the basis, and a cutoff point was determined accordingly. Accordingly, scores in the range of 1.00–3.22 from the F2 dimension were evaluated as low level, while scores in the range of 3.23–4.00 were evaluated as high level. According to this classification, high scores indicate that participants have a more positive and higher level of perception toward STREAM disciplines, while low scores show that this perception remains at a more limited level. In the interpretation of the scores related to the F3 dimension, a two-cluster structure was taken as the basis, and a cutoff point was determined accordingly. Accordingly, scores in the range of 1.00–2.80 from the F3 dimension were evaluated as low level, while scores in the range of 2.81–4.00 were evaluated as high level. These cut-off points have been interpreted by considering the 1.00–4.00 score range due to the 4-point Likert scale used for the rating. Therefore, the upper limit set for a high level in each sub-dimension has been taken as 4.00. According to this classification, high scores indicate that participants perceive the difficulties in implementing STREAM education at a higher level, while low scores show that this perception of difficulty is at a more limited level. These cut-off points should be considered sample-based and provisional. Therefore, they should be cross-validated in independent samples before being used for high-stakes decisions or definitive classification purposes.

Testing Measurement Invariance

To determine whether the measurement tool adapted to assess pre-service teachers' STREAM levels measured the same construct across female and male participants, multi-group confirmatory factor analysis was conducted to test measurement invariance. In evaluating measurement invariance, changes in fit indices were considered rather than relying solely on the chi-square difference test. For ΔCFI , a change of $\leq .01$ is commonly recommended as evidence of invariance (Cheung & Rensvold, 2002; French & Finch, 2006), while a more liberal criterion of $\leq .02$ has also been suggested for complex models and multiple-group comparisons (Rutkowski & Svetina, 2014). For $\Delta RMSEA$ and $\Delta SRMR$, changes of $\leq .015$ and $\leq .030$, respectively, have been recommended as acceptable thresholds (Chen, 2007; Rutkowski & Svetina, 2014). In the present study, ΔCFI , $\Delta RMSEA$, and $\Delta SRMR$ values were evaluated together when determining measurement invariance across gender.

Table 6
Measurement Invariance Results of the STREAM Scale by Gender

Model	χ^2	df	p	CFI	TLI	RMSEA	SRMR	ΔCFI	$\Delta RMSEA$	$\Delta SRMR$
Configural	835.047	592	< .001	.973	.970	.041	.077	—	—	—
Metric	899.287	615	< .001	.968	.967	.043	.079	-.005	.002	.002
Scalar	930.341	638	< .001	.967	.967	.043	.077	-.001	.000	-.002
Strict	973.562	664	< .001	.966	.966	.043	.080	-.001	.000	.003

Note. CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual. Δ values were calculated based on comparisons with the preceding model.

Table 6 presents the results of the measurement invariance of the STREAM Scale by gender. First, it was determined that the configural model tested showed acceptable/good fit, $\chi^2(592) = 835.047$, $p < .001$, CFI = .973, TLI = .970, RMSEA = .041, SRMR = .077. This finding indicates that the factor structure of the scale works similarly for female and male participants and that configural invariance was established. Then, a metric model where the factor loadings were constrained to be equal across groups was tested, and it was observed that the model fit continued to be at an acceptable level, $\chi^2(615) = 899.287$, $p < .001$, CFI = .968, TLI = .967, RMSEA = .043, SRMR = .079. When the change values between the configural model and the metric model were examined, they were found to be $\Delta CFI = -.005$, $\Delta RMSEA = .002$, and $\Delta SRMR = .002$. The fact that these changes remain below the recommended criteria indicates that factor loadings do not vary by gender and that metric invariance is achieved.

After the metric model, the scalar invariance model was tested, where the item thresholds/cut-off points were equalized between groups. It has been determined that the scalar model also has acceptable fit values, $\chi^2(638) = 930.341$, $p < .001$, CFI = .967, TLI = .967, RMSEA = .043, SRMR = .077. Compared to the metric model, $\Delta CFI = -.001$, $\Delta RMSEA = .000$, and $\Delta SRMR = -.002$ were calculated. This finding indicates that the initial levels of the scale items do not significantly differ by gender and that scalar invariance is achieved. Finally, the strict model, in which error variances are also equalized between groups, was examined. The fit indices obtained for the strict model are $\chi^2(664) = 973.562$, $p < .001$, CFI = .966, TLI = .966, RMSEA = .043, SRMR = .080, it has shown that the model demonstrates acceptable fit. When compared to the scalar model, it was found that $\Delta CFI = -.001$, $\Delta RMSEA = .000$, and $\Delta SRMR = .003$. The fact that these values remain within acceptable limits indicates that the error variances are also similar between gender groups and that strict invariance is achieved.

Table 7*One-Way ANOVA Results of F1, F2, and F3 Scores by Class Level and Department*

Variable	Group Variable	Groups	<i>n</i>	<i>M</i>	<i>SD</i>	Levene <i>p</i>	<i>F</i>	<i>p</i>	η^2	Sign. Diff.
F1	Grade Level	1	128	3.087	0.363	.714	0.443	.722	0.003	None
		2	116	3.037	0.355					
		3	144	3.058	0.392					
		4	110	3.082	0.398					
F2	Grade Level	1	128	3.264	0.417	.994	0.171	.916	0.001	None
		2	116	3.238	0.393					
		3	144	3.235	0.390					
		4	110	3.231	0.398					
F3	Grade Level	1	128	2.631	0.517	.539	2.231	.084	0.013	None
		2	116	2.699	0.491					
		3	144	2.769	0.560					
		4	110	2.781	0.509					
F1	Department	1	116	3.026	0.351	.058	6.687	<.001	0.064	6 > 1 2 > 5 3 > 5 4 > 5 6 > 5
		2	33	3.175	0.301					
		3	58	3.135	0.377					
		4	126	3.109	0.411					
		5	133	2.950	0.355					
		6	32	3.290	0.323					
F2	Department	1	116	3.193	0.383	.741	4.090	.001	0.040	2 > 5 4 > 5
		2	33	3.371	0.363					
		3	58	3.273	0.426					
		4	126	3.324	0.370					
		5	133	3.145	0.420					
		6	32	3.313	0.357					
F3	Department	1	116	2.607	0.560	.194	2.366	.039	0.023	3 > 1
		2	33	2.748	0.500					
		3	58	2.872	0.500					
		4	126	2.747	0.499					
		5	133	2.699	0.507					
		6	32	2.803	0.573					

Note. For grade level: 1 = first year, 2 = second year, 3 = third year, and 4 = fourth year. For department: 1 = Primary Mathematics Education, 2 = Chemistry Education, 3 = Science Education, 4 = Primary School Education, 5 = Turkish Language Education, and 6 = Art Education. SD = standard deviation. Since the Levene test results were not statistically significant, the assumption of homogeneity of variances was considered to be met. η^2 values were calculated based on the sum of squares. In the "Significant Difference" column, only statistically significant post hoc comparisons are presented.

According to the one-way ANOVA results, there was no significant difference in F1, F2, and F3 scores in terms of class level. Accordingly, it can be said that the students' F1, F2, and F3 scores are similar according to their grade levels. The non-significance of the Levene test results indicates that the assumption of homogeneity of variances is met.

In the analyses conducted according to the department variable, significant differences were found in the F1, F2, and F3 scores. F1 scores significantly varied according to the department variable ($F(5, 492) = 6.687, p < 0.001$). According to the post-hoc results, the F1 scores of the students in the Department of Art Education are significantly higher than those of the students in the Department of Primary School Mathematics Education and the Department of Turkish Language Education; the scores of the students in the Departments of Chemistry, Sci-

ence Education, and Classroom Teaching are significantly higher than those of the students in the Department of Turkish Language Education.

F2 scores also showed a significant difference according to the department variable ($F(5, 492) = 4.090$, $p = 0.001$). According to post-hoc results, the F2 scores of Chemistry Education and Classroom Teaching students are significantly higher than those of Turkish Language Education students.

A significant difference was also found in F3 scores according to the department variable ($F(5, 492) = 2.366$, $p = 0.039$). According to the post-hoc results, the F3 scores of the Science Education students are significantly higher than those of the Primary School Mathematics Education students. In general, while the class level variable does not have a significant effect on F1, F2, and F3 scores, the department variable is seen to create significant differences in all dimensions.

Discussion

The findings supported a three-factor structure for the Turkish version of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers. Perceptions of STREAM implementation, perceptions of STREAM disciplines, and perceived implementation challenges emerged as related but distinct dimensions. This structure is consistent with the multidimensional nature of STREAM education, which involves not only disciplinary integration but also pedagogical readiness, implementation competence, and contextual conditions (Johnston et al., 2022; Phang et al., 2023; Subramaniam et al., 2023; Sun & Zhong, 2024; Yang et al., 2023).

Confirmatory factor analysis showed an acceptable to good model fit after four items with low factor loadings and high error variances were removed. The item-removal decisions were based on statistical indicators, along with expert judgments on content coverage and theoretical relevance. This approach is consistent with contemporary test-adaptation standards, which emphasize that psychometric decisions should not rely solely on model fit but should also consider the construct represented by each item (AERA et al., 2014; Brown, 2015; DeVellis, 2017; ITC, 2017; Kline, 2016; Worthington & Whittaker, 2006). Two retained items had borderline loadings (.300 and .306) and relatively high error variances. Although their retention preserved the theoretical scope of the relevant factor, these items should be re-examined and, where necessary, revised in future validation studies using independent samples.

The reliability coefficients indicated acceptable internal consistency for all three subscales. The consistency of the findings across Cronbach's alpha, McDonald's omega, and composite reliability supports the reliability of the subscale scores, while avoiding reliance on a single coefficient that may be affected by restrictive assumptions such as tau-equivalence (Dunn et al., 2014; McNeish, 2018). Nevertheless, the convergent validity findings were more cautious. The AVE values for the first two factors were below the conventional .50 threshold, whereas composite reliability values exceeded .70, therefore provide acceptable evidence of construct reliability but limited evidence of convergent validity for these dimensions (Fornell & Larcker, 1981). By contrast, the HTMT values were below .85, supporting the discriminant validity of the three factors (Henseler et al., 2015). Taken together, these findings support interpretation at the subscale level rather than through a single total score.

The relationship between the first two STREAM dimensions was positive and moderate, indicating that participants who evaluated STREAM implementation positively also tended to value its disciplinary components. In contrast, perceived implementation challenges showed weak relationships with the other dimensions. This pattern suggests that recognizing the educational value of STREAM does not necessarily mean that pre-service teachers perceive its implementation as easy. They may hold positive views of STREAM while also anticipating barriers related to time, materials, professional knowledge, experience, and institutional support (Pakalnina & Grava, 2025; Phang et al., 2023). Similar distinctions between pedagogical beliefs, self-efficacy, and professional development needs have been reported in teacher education research (Chen et al., 2021; Yang et al., 2023).

The distinct position of the implementation-challenges factor also raises an important issue concerning the coherence of the overall instrument. Its low correlations and HTMT values with the first two factors indicate that it represents perceived barriers rather than a negative form of STREAM perception. The factor may therefore be most useful as a complementary diagnostic subscale that identifies contextual and implementation-related needs. Future research could test whether a bifactor model or a two-instrument structure better represents the distinction between positive STREAM perceptions and perceived implementation barriers. This interpretation is consistent with evidence that STREAM implementation challenges are often associated with contextual resources and professional support rather than attitudes alone (Pakalnina & Grava, 2025).

Criterion-related validity was supported by the positive relationships between the first two STREAM dimensions and engineering design self-efficacy. Pre-service teachers who perceived STREAM implementation and disciplines more positively also reported greater confidence in engineering design processes. This finding is theoretically consistent with Bandura's (1997) view that domain-specific self-efficacy influences individuals' willingness and persistence in performing related tasks. Engineering design is central to STREAM because it involves identifying problems, generating and evaluating solutions, developing products or prototypes, and integrating knowledge from multiple disciplines (Kahraman et al., 2024; Mumba et al., 2023; Ojeogwu & Mumba, 2025).

The findings also align with studies showing that engineering design and integrated STEM/STEAM experiences can strengthen pre-service teachers' design thinking, instructional competence, and science-engineering teaching self-efficacy (Hammack et al., 2024; Khamhaengpol et al., 2021; Lin et al., 2021; Ozkizilcik & Cebesoy, 2024). Similar associations between integrated STREAM experiences, implementation intentions, and teaching self-efficacy have been reported in different educational contexts (Menon et al., 2025; Subramaniam et al., 2023). These converging findings support selecting engineering design self-efficacy as a theoretically appropriate external criterion.

Perceived implementation challenges, however, were weakly or non-significantly related to engineering design self-efficacy. This result suggests that implementation difficulties may depend more strongly on external conditions than on individual confidence in engineering design. Limitations in time, resources, teacher preparation, infrastructure, and institutional support may remain even among individuals who feel competent in design-based activities (Ghazali et al., 2024; Phang et al., 2023). The weak relationship therefore reinforces the need to interpret implementation challenges separately from positive STREAM perceptions and individual self-efficacy.

The scale demonstrated configural, metric, scalar, and strict measurement invariance across gender. This finding indicates that the factor structure, loadings, thresholds, and error variances were comparable across female and male participants. Establishing measurement invariance is necessary for meaningful group comparisons and reduces the likelihood that observed differences reflect measurement bias rather than genuine variation in the construct (Chen, 2007; Cheung & Rensvold, 2002; Meredith, 1993; Putnick & Bornstein, 2016). However, because male participants represented a relatively small proportion of the sample, replication with more balanced groups remains necessary.

The findings are broadly consistent with international research showing that insufficient professional development and implementation resources can limit STREAM practices, whereas explicit exposure to integrated STEM/STEAM and engineering design can strengthen teacher self-efficacy and instructional readiness (Capobianco & Radloff, 2022; Ng et al., 2022; Phang et al., 2023). Nevertheless, cross-cultural validation is required before the Turkish form can be used confidently in other linguistic and educational contexts.

The empirical cut-off points derived for each subscale provide a preliminary framework for interpreting low, medium, and high score ranges. These classifications may support needs assessment, program evaluation, and professional development planning when subscale scores are reported separately. The third factor should be interpreted as an indicator of perceived implementation barriers rather than combined with the other dimensions in a total STREAM perception score. Because the thresholds were derived from the present sample, they should be treated as provisional and cross-validated before being used for consequential decisions. This subscale-based approach is appropriate for multidimensional instruments whose factors do not form a homogeneous higher-order structure (Hair et al., 2019).

No significant differences were observed across grade levels, suggesting that STREAM perceptions do not necessarily develop solely through progression in a teacher education program. More direct experiences, such as methods courses, engineering design activities, workshops, and practice-based learning, may be more influential than the number of years spent in the program (Hammack et al., 2024; Menon et al., 2025). In contrast, significant differences across academic departments indicate that disciplinary and pedagogical experiences may shape how pre-service teachers interpret STREAM implementation, disciplines, and challenges.

Departmental differences should not be interpreted as evidence that STREAM is more relevant to particular subject areas. Instead, they suggest that access to interdisciplinary and design-based experiences may vary across programs. Previous studies have shown that STEM modules and design-based activities can support conceptual development, creative thinking, and problem-solving across educational contexts (Arifin & Siew, 2023; Dedetürk et al., 2021). Teacher education programs should therefore integrate STREAM experiences not only into science and mathematics education but also into language, primary school, and arts education through activities suited to the epistemological and pedagogical characteristics of each discipline.

Limitations

This study has several limitations. Participants were selected through convenience sampling from a single national and institutional context. This limits the generalizability of the findings. The sample was also predominantly female. Most participants had limited prior exposure to STREAM-related training, coursework, or practical activities. Although measurement invariance across gender was supported, future studies should replicate the findings with more balanced, multi-site, and cross-cultural samples. These studies should also include participants with varying levels of STREAM experience.

The study relied on self-report measures. The responses may therefore have been influenced by social desirability or inaccurate self-assessment. Future research could combine scale scores with observations, performance-based tasks, lesson-plan analyses, interviews, or focus groups. This would provide a more comprehensive account of STREAM perceptions, implementation challenges, and engineering design self-efficacy.

Evidence of convergent validity was limited for the first two factors. Their AVEs were below the conventional threshold. In addition, two retained items had borderline factor loadings and relatively high error variances. These items and the factor structure should be re-examined in independent samples. Item wording should be revised where necessary. The empirical cut-off points should be considered provisional until they are cross-validated in different samples.

Finally, the cross-sectional design did not allow for examining changes in STREAM perceptions or engineering design self-efficacy over time. Longitudinal and experimental studies are needed to determine how STREAM-based courses, workshops, and practice-oriented interventions influence these constructs.

Conclusions and Implications

This study provided psychometric evidence for the Turkish adaptation of the STREAM Perception and Implementation Challenges Scale for Pre-Service Teachers. Findings supported a 26-item, three-factor structure: perceptions of STREAM implementation, perceptions of STREAM disciplines, and perceived implementation challenges. The subscales showed acceptable internal consistency. Measurement invariance across gender showed that the instrument functions similarly for female and male pre-service teachers.

Perceptions of STREAM implementation and STREAM disciplines were positively associated with engineering design self-efficacy. In contrast, perceived implementation challenges showed only weak or non-significant relationships with this construct. The low associations between the third factor and the other dimensions indicate that STREAM-PICS scores should be interpreted separately at the subscale level rather than combined into a single total score. Although construct reliability and discriminant validity were supported, the low AVE for the first two factors indicates that convergent validity remains limited.

The scale may be used in teacher education programs to identify pre-service teachers' perceptions of STREAM education, their evaluations of STREAM disciplines, and the barriers they anticipate during implementation. By administering the scale, programs can better assess specific needs, evaluate current curricula, and plan targeted professional development activities. High scores on the implementation-challenges subscale specifically signal areas where more support may help teacher educators identify needs related to practical experience, instructional resources, interdisciplinary planning, or implementation support. The empirical cut-off points aid initial interpretation of subscale scores, but they should be used cautiously for decision-making until confirmed in independent samples.

The absence of grade-level differences suggests that STREAM-related perceptions may not develop solely through progression in a teacher education program. In contrast, differences across academic departments indicate that disciplinary and program-related experiences may shape these perceptions. These findings imply that teacher education programs should provide structured, practice-oriented STREAM experiences across all areas- science, mathematics, language, primary school, and arts- rather than confining them to particular subjects. Future research should validate the two items with borderline factor loadings, test the factor structure and cut-off in more diverse samples, and evaluate how STREAM-based courses and interventions influence perceptions through longitudinal or experimental designs.



Acknowledgements

This article is derived from the M.Sc. thesis by the first author under the supervision of the second author. The authors would like to thank the participating pre-service teachers for their valuable time and cooperation.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and relevant national regulations. Ethical approval for the study was obtained from the Social and Humanities Ethics Committee Committee], [Çanakkale Onsekiz Mart University] (Approval No: [E-80110942-050.99-2500208162 - 2025-YÖNP-0709 -55/72], Date: [22.08.2025]).

References

- Alanazi, F. H., & Al-Zahrani, E. (2025). Examination of teachers' self-effectiveness toward teaching integrated science, technology, reading, engineering, art, and mathematics: A cross-sectional study. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(3), Article em2606. <https://doi.org/10.29333/ejmste/16071>
- Alghamdi, A. A. (2023). Exploring early childhood teachers' beliefs about STEAM education in Saudi Arabia. *Early Childhood Education Journal*, 51(2), 247–256. <https://doi.org/10.1007/s10643-021-01303-0>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. <https://www.apa.org/ethics/code/>
- Arifin, S., & Siew, N. M. (2023). Integration of socioscientific approach and design thinking: An entrepreneurial creative thinking module for STEM education. *Journal of Baltic Science Education*, 22(5), 767–780. <https://doi.org/10.33225/jbse/23.22.767>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.
- Belbase, S., Mainali, B. R., Kasemsukpipat, W., Tairab, H., Gochoo, M., & Jarrah, A. (2022). At the dawn of science, technology, engineering, arts, and mathematics (STEAM) education: Prospects, priorities, processes, and problems. *International Journal of Mathematical Education in Science and Technology*, 53(11), 2919–2955. <https://doi.org/10.1080/0020739X.2021.1922943>
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258. <https://doi.org/10.1177/0049124192021002005>
- Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş., & Demirel, F. (2019). *Eğitimde bilimsel araştırma yöntemleri* [Scientific research methods in education] (26. baskı). Pegem Akademi.
- Capobianco, B. M., & Radloff, J. (2022). Elementary preservice teachers' trajectories for appropriating engineering design-based science teaching. *Research in Science Education*, 52(5), 1623–1641. <https://doi.org/10.1007/s11165-021-10020-y>
- Capobianco, B. M., Radloff, J., & Clingerman, J. (2022). Facilitating preservice elementary science teachers' shift from learner to teacher of engineering design-based science teaching. *International Journal of Science and Mathematics Education*, 20(4), 747–767. <https://doi.org/10.1007/s10763-021-10193-y>
- Chen, F. F. (2007). Sensitivity of goodness-of-fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Chen, Y.-L., Huang, L.-F., & Wu, P.-C. (2021). Preservice preschool teachers' self-efficacy in and need for STEM education professional development: STEM pedagogical belief as a mediator. *Early Childhood Education Journal*, 49(2), 137–147. <https://doi.org/10.1007/s10643-020-01055-3>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE.
- Dedetürk, A., Kirmızıgül, A. S., & Kaya, H. (2021). The effects of STEM activities on 6th grade students' conceptual development of sound. *Journal of Baltic Science Education*, 20(1), 21–37. <https://doi.org/10.33225/jbse/21.20.21>
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). SAGE.
- Dinh, D. T., Fujinami, T., & Huynh, V. N. (2019). Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In J. Chen, V. Huynh, G. N. Nguyen, & X. Tang (Eds.), *Knowledge and systems sciences: KSS 2019* (Communications in Computer and Information Science, Vol. 1103). Springer. https://doi.org/10.1007/978-981-15-1209-4_1
- Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.2307/3151312>



- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3
- Ghazali, A., Mohamad Ashari, Z., Hardman, J., & Abu Yazid, A. (2024). Development and effectiveness of the E-Sky module based on PBL in the teaching and facilitation process of early science. *Journal of Baltic Science Education*, 23(2), 221–239. <https://doi.org/10.33225/jbse/24.23.221>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
- Hammack, R., Yeter, I. H., Pavlovich, C., & Boz, T. (2024). Pre-service elementary teachers' science and engineering teaching self-efficacy and outcome expectancy: Exploring the impacts of efficacy source experiences through varying course modalities. *International Journal of STEM Education*, 11, Article 4. <https://doi.org/10.1186/s40594-024-00464-9>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). International Test Commission. <https://www.intestcom.org>
- Johnston, K., Kervin, L., & Wyeth, P. (2022). STEM, STEAM and makerspaces in early childhood: A scoping review. *Sustainability*, 14(20), Article 13533. <https://doi.org/10.3390/su142013533>
- Kahraman, N., Yecan, E., Taşkın Ekici, F., & Kara, İ. (2024). Mühendislik Tasarım Becerileri Özyeterlik Algı Ölçeği'nin (MTBÖAÖ) geliştirilmesi. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 61, 458–480. <https://doi.org/10.9779/pauefd.1284927>
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley. <https://doi.org/10.1002/9780470316801>
- Khamaengpol, A., Sriprom, M., & Chuamchaitrakool, P. (2021). Development of STEAM activity on nanotechnology to determine basic science process skills and engineering design process for high school students. *Thinking Skills and Creativity*, 39, Article 100796. <https://doi.org/10.1016/j.tsc.2021.100796>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Laksmiwati, P. A., Lavicza, Z., Cahyono, A. N., Alagic, M., & Mumcu, F. (2024). When engineering design meets STEAM education in hybrid learning environment: Teachers' innovation key through design heuristics. *Asia Pacific Journal of Education*. Advance online publication. <https://doi.org/10.1080/02188791.2024.2373226>
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, 48(3), 936–949. <https://doi.org/10.3758/s13428-015-0619-7>
- Lin, K.-Y., Wu, Y.-T., Hsu, Y.-T., & Williams, P. J. (2021). Effects of infusing the engineering design process into STEM project-based learning to develop preservice technology teachers' engineering design thinking. *International Journal of STEM Education*, 8, Article 1. <https://doi.org/10.1186/s40594-020-00258-9>
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130–149. <https://doi.org/10.1037/1082-989X.1.2.130>
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, 23(3), 412–433. <https://doi.org/10.1037/met0000144>
- Menon, D., Wieselmann, J. R., Haines, S., Asim, S., Koch, A., & Cox, D. (2025). Preservice elementary teachers' integrated STEM teaching self-efficacy: Contributing sources within STEM education courses. *AERA Open*, 11(1), 1–20. <https://doi.org/10.1177/23328584251321472>
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543. <https://doi.org/10.1007/BF02294825>
- Mumba, F., Rutt, A., & Chabalengula, V. M. (2023). Representation of science and engineering practices and design skills in engineering design-integrated science units developed by pre-service teachers. *International Journal of Science and Mathematics Education*, 21(2), 439–461. <https://doi.org/10.1007/s10763-022-10266-6>
- Ng, A., Kewalramani, S., & Kidman, G. (2022). Integrating and navigating STEAM (inSTEAM) in early childhood education: An integrative review and inSTEAM conceptual framework. *Eurasia Journal of Mathematics, Science and Technology Education*, 18(7), Article em2133. <https://doi.org/10.29333/ejmste/12174>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Ojeogwu, J., & Mumba, F. (2025). Preparing pre-service teachers in engineering design integrated science teaching: A systematic review. *Journal of Science Education and Technology*, 35(1), 237–260. <https://doi.org/10.1007/s10956-025-10247-8>
- Ozkizilcik, M., & Cebesoy, U. B. (2024). The influence of an engineering design-based STEM course on pre-service science teachers' understanding of STEM disciplines and engineering design process. *International Journal of Technology and Design Education*, 34(2), 727–758. <https://doi.org/10.1007/s10798-023-09837-7>
- Pakalnina, R., & Grava, J. (2025). Challenges for preschool teachers in STREAM education. *EURASIA Journal of Mathematics, Science and Technology Education*, 21(11), Article em2726. <https://doi.org/10.29333/ejmste/17288>
- Phang, W. L. A., Chong, M. C., Che Mustafa, M., & Mohd Jamil, M. R. (2023). Issues and challenges for the implementation of preschool STREAM education: What do the preschool teachers say? *Southeast Asia Early Childhood Journal*, 12(1), 54–69. <https://doi.org/10.37134/saecj.vol12.1.5.2023>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>

- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. <https://doi.org/10.1037/a0029315>
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74(1), 31–57. <https://doi.org/10.1177/0013164413498257>
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Subramaniam, V., Karpudewan, M., & Roth, W. M. (2023). Unveiling the teachers' perceived self-efficacy to practice integrated STrEaM teaching. *The Asia-Pacific Education Researcher*, 32(3), 327–337. <https://doi.org/10.1007/s40299-022-00655-4>
- Sun, W., & Zhong, B. (2024). Integrating reading and writing with STEAM/STEM: A systematic review on STREAM education. *Journal of Engineering Education*, 113(4), 939–958. <https://doi.org/10.1002/jee.20569>
- Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
- Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, 73(6), 913–934. <https://doi.org/10.1177/0013164413495237>
- Whittington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838. <https://doi.org/10.1177/0011000006288127>
- Yang, W., Wu, R., & Li, J. (2023). Development and validation of the STEM Teaching Self-efficacy Scale (STSS) for early childhood teachers. *Current Psychology*, 42, 7275–7283. <https://doi.org/10.1007/s12144-021-02074-y>

Received: June 12, 2026

Revised: July 12, 2026

Accepted: August 17, 2026

Cite as: Yerli, H., & Girgin, D. (2026). Perceptions of stream and self-efficacy in engineering design skills among pre-service teachers: Scale adaptation, validity-reliability, and relational evidence. *Journal of Baltic Science Education*, 25(4), 631–649. <https://doi.org/10.33225/jbse/26.25.631>



Halime Yerli

MSc, School of Graduate Studies, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye.
E-mail: halime17yerli@gmail.com
ORCID: <https://orcid.org/0009-0007-8833-4206>

Derya Girgin
(Corresponding author)

PhD, Associate Professor, Department of Education Sciences, Faculty of Education, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye,
E-mail: deryagirgin@comu.edu.tr
ORCID: <https://orcid.org/0000-0002-6114-7925>

